Indexing Biotic Interactions in GBIF data

José Augusto Salim[‡,§], Katja Chantre Seltmann[|], Jorrit H Poelen[|,¶], Antonio Mauro Saraiva[‡]

‡ Universidade de São Paulo, Escola Politécnica, São Paulo, Brazil
§ Universidade Estadual de Campinas, Instituto de Biologia, Depto. de Biologia Vegetal, Campinas, Brazil
| Cheadle Center for Biodiversity and Ecological Restoration, University of California - Santa Barbara, Santa Barbara, United States of America
¶ Ronin Institute, Montclair NJ, United States of America

Corresponding author: José Augusto Salim (joseasalim@usp.br)

Abstract

The Global Biodiversity Information Facility (GBIF 2022a) has indexed more than 2 billion occurrence records from 70,147 datasets. These datasets often include "hidden" biotic interaction data because biodiversity communities use the Darwin Core standard (DwC, Wieczorek et al. 2012) in different ways to document biotic interactions. In this study, we extracted biotic interactions from GBIF data using an approach similar to that employed in the Global Biotic Interactions (GloBI; Poelen et al. 2014) and summarized the results. Here we aim to present an estimation of the interaction data available in GBIF, showing that biotic interaction claims can be automatically found and extracted from GBIF. Our results suggest that much can be gained by an increased focus on development of tools that help to index and curate biotic interaction data in existing datasets. Combined with data standardization and best practices for sharing biotic interactions, such as the initiative on plant-pollinators interaction (Salim 2022), this approach can rapidly contribute to and meet open data principles (Wilkinson 2016).

We used Preston (Elliott et al. 2020), open-source software that versions biodiversity datasets, to copy all GBIF-indexed datasets. The biodiversity data graph version (Poelen 2020) of the GBIF-indexed datasets used during this study contains 58,504 datasets in Darwin Core Archive (DwC-A) format, totaling 574,715,196 records. After retrieval and verification, the datasets were processed using Elton. Elton extracts biotic interaction data and supports 20+ existing file formats, including various types of data elements in DwC records. Elton also helps align interaction claims (e.g., host of, parasite of, associated with) to the Relations Ontology (RO, Mungall 2022), making it easier to discover datasets across a heterogeneous collection of datasets. Using specific mapping between interaction claims found in the DwC records to the terms in RO*[1], Elton found 30,167,984 *potential records* (with non-empty values for the scanned DwC terms) and 15,248,478 records with recognized interaction types.

Taxonomic name validation was performed using Nomer, which maps input names to names found in a variety of taxonomic catalogs. We only considered an interaction

record valid where the interaction type could be mapped to a term in RO and where Nomer found a valid name for *source* and *target* taxa.

Based on the workflow described in Fig. 1, we found 7,947,822 interaction records (52% of the *potential* interactions). Most of them were generic interactions (*interacts_with*, 87.5%), but the remaining 12.5% (993,477 records) included host-parasite and plant-animal interactions. The majority of the interactions records found involved plants (78%), animals (14%) and fungi (6%).

In conclusion, there are many biotic interactions embedded in existing datasets registered in large biodiversity data indexers and aggregators like iDigBio, GBIF, and BioCASE. We exposed these biotic interaction claims using the combined functionality of biodiversity data tools Elton (for interaction data extraction), Preston (for reliable dataset tracking) and Nomer (for taxonomic name alignment). Nonetheless, the development of new vocabularies, standards and best practice guides would facilitate aggregation of interaction data, including the diversification of the GBIF data model (GBIF 2022b) for sharing biodiversity data beyond occurrences data. That is the aim of the TDWG Interest Group on Biological Interactions Data (TDWG 2022).

Keywords

global biotic interactions, Preston, Elton, Nomer, Darwin Core

Presenting author

José Augusto Salim

Presented at

TDWG 2022

Conflicts of interest

References

- Elliott M, Poelen J, Fortes JB (2020) Toward reliable biodiversity dataset references. Ecological Informatics 59 https://doi.org/10.1016/j.ecoinf.2020.101132
- GBIF (2022a) What is GBIF? URL: https://www.gbif.org/what-is-gbif
- GBIF (2022b) Diversifying the GBIF Data Model. https://www.gbif.org/composition/HjlTr705BctcnaZkcjRJq/data-model. Accessed on: 2022-6-29.
- Mungall C, et al. (2022) oborel/obo-relations: v2022-05-23. Zenodo https://doi.org/10.5281/zenodo.593101

- Poelen J, Simons J, Mungall C (2014) Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. Ecological Informatics 24: 148-159. https://doi.org/10.1016/j.ecoinf.2014.08.005
- Poelen J (2020) A biodiversity dataset graph: GBIF, iDigBio, BioCASe hash://sha256/8aacce08462b87a345d271081783bdd999663ef90099212c8831db399fc0831b. Zenodo https://doi.org/10.5281/zenodo.1472393
- Salim J, et al. (2022) Data standardization of plant-pollinator interactions. GigaScience 11 https://doi.org/10.1093/gigascience/giac043
- TDWG (2022) Biological Interactions Data Interest Group. https://www.tdwg.org/community/interaction/. Accessed on: 2022-6-30.
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D, et al. (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7 (1). https://doi.org/10.1371/journal.pone.0029715
- Wilkinson M, et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18

Endnotes

[*1] https://raw.githubusercontent.com/globalbioticinteractions/prestonocene/main/mapping_unsupported_interactions.tsv

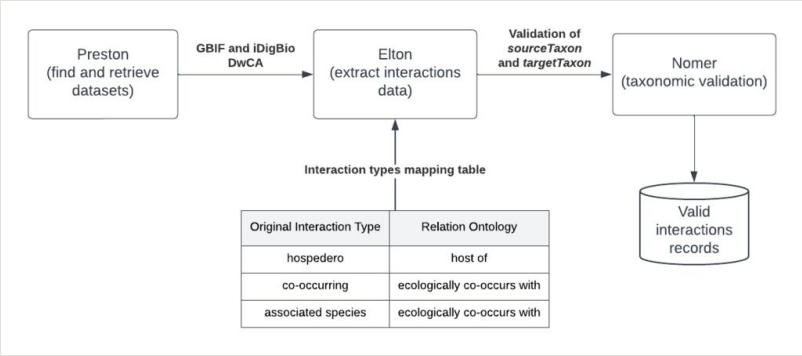

Figure 1.

Preston, Elton, Nomer workflows to retrieve and process biotic interactions from GBIF data.